МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Український державний морський технічний університет
імені адмірала Макарова

МЕТОДИЧНІ ВКАЗІВКИ ДО ВИВЧЕННЯ КУРСУ "МАТЕМАТИЧНА СТАТИСТИКА ДЛЯ ЛІНГВІСТІВ"

Рекомендовано Методичною радою УДМТУ

Миколаїв 2002

УДК 517: 618.3

Кузнецов А.М. Методичні вказівки до вивчення курсу "Математична статистика для лінгвістів". – Миколаїв: УДМТУ, 2002. – 28 с.

***Кафедра** вищої математики*

Методичні вказівки складені на основі лекцій для лінгвістів з математичної статистики. Цей курс є логічним продовженням лекцій з теорії імовірностей. Викладання теоретичного матеріалу супроводжується прикладами розв'язання задач.

Рекомендовано для студентів-лінгвістів.

Рецензент канд. техн. наук, доцент Є.Ю. Нєдєлько.

І. ПРЕДМЕТ МАТЕМАТИЧНОЇ СТАТИСТИКИ

Ми вважаємо, що студент вже обізнаний з основними поняттями теорії імовірностей, а саме: випадковою подією, випадковою величиною, математичним сподіванням (м. с.), дисперсією, законами розподілу у вигляді ряду розподілу, функціями розподілу, щільностями розподілу (нормальним законом розподілу).

В теорії імовірностей ми могли заздалегідь вважати відомою, наприклад, ймовірність якої-небудь випадкової події або функцію розподілу випадкової величини. Але на практиці частіше буває інакше: ми ставимо дослід і вказані характеристики одержуємо з нього.

Наука, яка займається розробкою методів збирання і обробки дослідних даних з метою вивчення закономірностей масових випадкових явищ, називається *математичною статистикою*.

В основі математичної статистики лежить ряд вихідних понять, без попереднього знайомства з якими неможливе вивчення методів обробки дослідних даних. Зупинимося на них.

ІІ. ГЕНЕРАЛЬНА СУКУПНІСТЬ І ВИБІРКА

Нехай треба дослідити яку-небудь якісну чи кількісну ознаку, властиву великій групі однорідних об'єктів. Наприклад, перевірити на стандартність чи нестандартність виготовлених деталей, розмір деталей, бюджет сім'ї, читацький попит на художню літературу, якість знань випускників з української мови і т.п. Найчастіше суцільне обстеження однорідних об'єктів не проводиться, бо це або

фізично неможливо, або економічно невигідно. Тому за результатами вивчення невеликої частини об'єктів одержують з достатньою для практики вірогідністю необхідну інформацію про всю сукупність. Такий метод дослідження носить назву *вибіркового*.

Усі об'єкти, що вивчаються, називаються *генеральною сукупністю*. Її величина N – об'єм генеральної сукупності.

Частина об'єктів, випадково відібрана із генеральної сукупності для перевірки, називається *вибірковою сукупністю* або *вибіркою*.

Її величина n ($n \le N$) – об'єм вибірки.

В основному розрізняють два способи вибору об'єктів з генеральної сукупності – *випадковий* і *невипадковий*. До останнього відноситься серійний добір, за яким об'єкти добираються не по одному, а серією.

У свою чергу, випадковий добір може бути добором з поверненням і без нього.

Випадковий добір об'єктів з генеральної сукупності називається *добором* (*вибіркою*) *з поверненням*, якщо перед випадковим добором наступного об'єкта раніше відібраний об'єкт повертається до генеральної сукупності.

Випадковий добір об'єктів з генеральної сукупності називається *добором* (*вибіркою*) *без повернення*, якщо перед випадковим добором наступного об'єкта раніше відібрані випадково об'єкти не повертаються до генеральної сукупності.

Вибірка з поверненням має перевагу з точки зору методики вивчення властивостей генеральної сукупності, тому що в цьому випадку проводять повторні незалежні випробування.

При достатньо великому об'ємі генеральної сукупності оцінки її характеристик, здобуті на основі вибірки без повернення, істотно не відрізняються від оцінок, здобутих на основі вибірки з поверненням.

На практиці частіше користуються вибіркою без повернення.

Характеристики вибірки приймаються як наближені значення відповідних характеристик генеральної сукупності.

Доведено, що якщо кожний об'єкт вибірки вибраний випадково, то вибірка дає найбільш повну інформацію про генеральну сукупність. Тоді говорять, що вибірка *репрезентативна* (представницька).

III. ЧИСЛОВІ ХАРАКТЕРИСТИКИ СТАТИСТИЧНОГО РОЗПОДІЛУ У ЛІНГВІСТИЦІ

Числові характеристики служать для математичної оцінки результатів спостережень або досліджень.

До числа елементарних інструментів спостереження за дією статистичних законів відносять *частоту*, *відносну частоту* і *відхилення* від середньої частоти.

Частотою якої-небудь події називають число її появ у відрізку дійсності, який ми спостерігаємо. Для лінгвістів таким відрізком може бути текст того чи іншого об'єму, тої чи іншої довжини. *Відносною частотою* називаємо відношення частоти до об'єму вибірки.

1. Обчислення вибіркової середньої

Мовознавець, що користується в дослідженнях статистичними методами, робить свої висновки завдяки результатам підрахунку частот явищ, що вивчаються у творах письменників, науковців тощо.

Нехай вивчається деяка випадкова величина X, а x_i – її частота у i-й вибірці. Результати вимірів представимо у вигляді таблиці, в першому рядку якої вказується номер виміру i, а в другому – результат виміру x_i:

i	1	2	3	….	k
x_i	x_1	x_2	x_3	….	x_k

Цю таблицю в математичній статистиці називають *статистичним рядом*.

Вибіркова середня $\bar{x}$ у даному випадку підраховується за формулою

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i}{k}. \qquad (1)$$

Приклад. Вивчалася частота іменників у повісті І. Франка "Захар Беркут".

Взято 10 вибірок, кожна по 500 слів. Одержали наступний ста-

тистичний ряд:

i	1	2	3	4	5	6	7	8	9	10
x_i	115	111	110	127	105	113	123	121	126	122

Знайти вибіркову середню $\bar{x}$.

Розв'язання:

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{115+111+...+126+122}{10} = \frac{1173}{10} = 117,3 \approx 117.$$

Якщо деякі частоти x_i повторюються і m_i – число таких повторень, m – число різних x_i, то вибіркова середня обчислюється за формулою

$$\bar{x} = \frac{\sum_{i=1}^{m} m_i x_i}{k} \qquad \left(k = \sum_{i=1}^{m} m_i \right). \tag{2}$$

Приклад. Вивчається частота прикметників у прозі К. Федіна. Взято 10 вибірок із авторської художньої прози по 500 слів кожна. Результати записані у таблицю:

x_i	43	59	60	69	71	72	73	83
m_i	1	1	1	2	2	1	1	1

$$k = \sum_{i=1}^{8} m_i = 10.$$

Знайти вибіркову середню $\bar{x}$.

Розв'язання:

$$\bar{x} = \frac{\sum_{i=1}^{8} m_i x_i}{10} = \frac{1 \cdot 43 + 1 \cdot 59 + 1 \cdot 60 + 2 \cdot 69 + 2 \cdot 71 + ... + 1 \cdot 83}{10} = 67.$$

Формули (1) і (2) використовуються тільки у випадку вибірок однакового об'єму.

Пояснимо на прикладі, що робити, якщо дані одержані із вибірок різного об'єму.

Приклад. У таблиці вказані об'єми n вибірок (кількість слів) і відповідні частоти x_i ознаки, що вивчається:

n	600	120	1000	1200	800
x_i	5	1	8	8	4

Треба встановити, яка вибіркова середня для вибірки із 1000 слів.

Розв'язання. Вибіркова відносна частота у кожній вибірці відповідно дорівнює $\frac{5}{600}, \frac{1}{120}, \frac{8}{1000}, \frac{8}{1200}, \frac{4}{800}$. Середня відносна частота з врахуванням того, що вибірок було 5: $\frac{1}{5}\left(\frac{5}{600}+\frac{1}{120}+\frac{8}{1000}+\right.$ $\left.+\frac{8}{1200}+\frac{4}{800}\right)$. Тоді вибіркова середня для 1000 слововживань $\bar{x}=\frac{1000}{5}\left(\frac{5}{600}+\frac{1}{120}+\frac{8}{1000}+\frac{8}{1200}+\frac{4}{800}\right)\approx 7.$

Взагалі, якщо $x_1, x_2, ..., x_k$ – частоти, $n_1, n_2, ..., n_k$ – відповідні об'єми вибірок, n – деякий об'єм, для якого підраховується вибіркова середня, а k – число вибірок, то

$$\bar{x}=\frac{n}{k}\sum_{i=1}^{k}\frac{x_i}{n_i}. \qquad (3)$$

Цей приклад показує, наскільки стає складнішою обчислювальна робота при вибірках нерівного об'єму. Особливо небезпечно екстраполювати частоту, яку спостерігаємо для малої вибірки, на частоту більшого об'єму.

Наприклад, в англійському технічному тексті із 100 словосполучень дієприкметник зустрічається 2–3 рази. Ясно, що таку частоту на уривок з більшою довжиною екстраполювати не можна, бо згідно правила пропорційності у вибірці з 1000 словосполучень ця ознака повинна б повторюватися 20–30 разів! На практиці ж він зустрічається приблизно 9 разів. Тому при нерівних вибірках краще користуватися не середньою арифметичною, а долею (відносною частиною).

2. Стандартне відхилення

Середня є дуже важливим параметром, що характеризує випадкову величину (в. в.), але знання тільки однієї середньої не дає повної уяви про поведінку в. в.

Наприклад, в. в. X приймає один раз значення 0, 20, 40, а другий раз 19, 20, 21. Їх середня $\bar{x}$ = 20. Але можливі значення x_i порізному сконцентровані біля своєї середньої $\bar{x}$.

Позначимо через δ_i різницю $(x_i - \bar{x})$ і назвемо її *відхиленням*. Але дослідника найчастіше цікавить середнє відхилення. Для цього спочатку обчислимо дисперсію:

$$\sigma^2 = \frac{\sum_{i=1}^{k} \delta_i^2}{k} = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2}{k},$$

а потім середньоквадратичне або стандартне відхилення:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k} (x_i - \bar{x})^2}{k}}. \qquad (4)$$

Якщо у формулі (4) замінити k на $k - 1$, то одержимо так звану *незсунену* оцінку середньоквадратичного відхилення, яка позначається через S:

$$S = \sqrt{\frac{\sum_{i=1}^{k} (x_i - \bar{x})^2}{k-1}}. \qquad (5)$$

Якщо k велике, то різниці при обчисленні за формулами (4) і (5) майже немає.

Приклад. Для 5 вибірок, кожна з яких складається з 500 слів, одержані наступні частоти дієслів (табл. 1).

Знайти стандартне відхилення.

Розв'язання:

$$\sigma = \sqrt{\frac{166}{5}} = \sqrt{33{,}2} \approx 5{,}76; \qquad S = \sqrt{\frac{166}{4}} = \sqrt{41{,}5} \approx 6{,}44.$$

На практиці замість формули (5) користуються іншою формулою. Перетворимо вираз $\sum_{i=1}^{k}(x_i-\bar{x})^2 = \sum_{i=1}^{k}(x_i{}^2-2x_i\bar{x}-\bar{x}^2) = \sum_{i=1}^{k}x_i{}^2 - 2\bar{x}\cdot k\frac{\sum_{i=1}^{k}x_i}{k}+\sum_{i=1}^{k}\bar{x}^2 = \sum_{i=1}^{k}x_i^2 - 2\bar{x}^2\cdot k + k\cdot\bar{x}^2 = \sum_{i=1}^{k}x_i^2 - k\cdot\bar{x}^2.$

Тоді

$$S=\sqrt{\frac{\sum_{i=1}^{k}x_i{}^2-k\bar{x}^2}{k-1}}. \qquad (6)$$

Таблиця 1. **Частоти дієслів**

Вибірки	x_i	$x_i-\bar{x}$	$(x_i-\bar{x})^2$	x_i^2
1	95	–1	1	9025
2	87	–9	81	7568
3	94	–2	4	8836
4	104	+8	64	10816
5	100	+4	16	10000
Σ	$\bar{x}=96$		166	46246

І для даного прикладу

$$S=\sqrt{\frac{46246-5\cdot 96^2}{4}}=6{,}44.$$

Зауважимо, що якщо m_i – число повторень x_i, то

$$S=\sqrt{\frac{\sum_{i=1}^{m}m_i(x_i-\bar{x})^2}{k-1}} \quad \text{або} \quad S=\sqrt{\frac{\sum_{i=1}^{m}m_i x_i{}^2-k\bar{x}^2}{k-1}}.$$

Введемо ще одне поняття математичної статистики – *ймовірна помилка при визначенні середньої частоти* (помилка середньої арифметичної). Справа у тому, що наші вибіркові дані не дають нам повної інформації про ту дійсну середню, яка характеризує усю генеральну сукупність. Наприклад, якщо на основі 20 вибірок ми одержали середню частоту дієслова у текстах Шевченка 110 одиниць, то це ще не означає, що "дійсна середня" всіх текстів Шевченка, із яких бралися вибірки, дорівнює 110. Цю дійсну середню ми не знаємо. Але щоб мати про неї наближену уяву, ми і визначаємо вибіркову середню частоту. Дійсна середня повинна бути десь біля нашої середньої. Але де саме? В якому інтервалі частот? Для відповіді на це питання і використовується знання ймовірної помилки в означенні середньої. Позначимо її через Δ. Але середня $\bar{x}$ коливається біля математичного сподівання (м. с.) m_x випадкової величини X,

тобто

$$\bar{x} - \Delta < m_x < \bar{x} + \Delta .$$

Одержали так званий довірчий (надійний) інтервал ($\bar{x} - \Delta$, $\bar{x} + \Delta$), у який попадає м.с. ознаки X, що вивчається, з надійністю β (надійна ймовірність).

Доведено, що ймовірна похибка при визначенні середньої обчислюється за формулою

$$\Delta = \frac{ts}{\sqrt{k}} \quad \text{або} \quad \Delta = \frac{t\sigma}{\sqrt{k}},$$

де t – особливий коефіцієнт, який залежить від числа вибірок k і обчислюється за табл. 2. Крім того, він залежить від β (як правило, $\beta = 0{,}95; 0{,}92; 0{,}9, \ldots$). Визначимо, наприклад, довірчий інтервал для прикладу, поданому вище: $\bar{x} = 96$; $k = 5$; $S = 6{,}44$.

Таблиця 2. **Значення *t***

k–1	β					
	0,8	0,9	0,95	0,98	0,99	0,999
1	3,08	6,31	12,71	31,8	63,7	63,7
2	1,886	2,92	4,30	6,96	9,92	31,6
3	1,638	2,35	3,18	4,54	5,84	12,94
4	1,533	2,13	2,78	3,75	4,60	8,61
5	1,476	2,02	2,57	3,36	4,03	6,86
6	1,440	1,943	2,45	3,14	3,71	5,96
7	1,415	1,895	2,36	3,00	3,50	5,40
8	1,397	1,860	2,31	2,90	3,36	5,04
9	1,383	1,833	2,26	2,82	3,25	4,78
10	1,372	1,812	2,23	2,76	3,17	4,59
11	1,363	1,796	2,20	2,72	3,11	4,99
12	1,356	1,782	2,18	2,68	3,06	4,32
13	1,350	1,771	2,16	2,65	3,01	4,22
14	1,345	1,761	2,14	2,62	2,98	4,14
15	1,341	1,753	2,13	2,60	2,95	4,07
16	1,337	1,746	2,12	2,58	2,92	4,02
17	1,333	1,740	2,11	2,57	2,90	3,96
18	1,330	1,734	2,10	2,55	2,88	3,92
19	1,328	1,729	2,09	2,54	2,86	3,88
20	1,325	1,725	2,09	2,53	2,84	3,85

Візьмемо $\beta = 0{,}95$ (95 %), тоді згідно з табл. 2 $t = 2{,}78$.

$$\Delta = \frac{2{,}78 \cdot 6{,}44}{\sqrt{5}} = \frac{17{,}9}{2{,}34} \approx 7{,}65 \approx 8;\quad 96 - 8 < m_x < 96 + 8 \text{ або } 88 < m_x < 104,$$

тобто з імовірністю у 0,95 інтервал (88;104) накриває невідоме математичне сподівання m_x.

3. Статистична оцінка розходжень між вибірковими частотами

Припустимо, що 10 текстових вибірок по 500 слів кожна дали такий ряд розподілу:

i	1	2	3	4	5	6	7	8	9	10
x_i	98	87	102	105	123	108	85	78	110	104

Вибіркова середня частота $\bar{x} = 100$.

Постає питання: чи одержані вони при одній і тій же самій імовірності, при одному і тому ж законі розподілу, наприклад нормальному? Якщо так, то ці розбіжності випадкові, пов'язані з обмеженою кількістю спостережень і, отже, статистично закономірні. Або, може ці відхилення від середньої частоти виникли внаслідок порушення статистичного закону, внаслідок зміни ймовірності протягом досліду. Якщо так, то ці коливання частот не випадкові, вони істотні.

Як встановити, випадкові чи ні відхилення вибіркових частот від їх середньої? В математичній статистиці одним із таких критеріїв служить критерій згоди "хі – квадрат" (критерій згоди Пірсона). У разі вибірок однакової довжини

$$\chi^2 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2}{\bar{x}}. \qquad (6)$$

Назва "критерій згоди" виникла тому, що він звіряє, узгоджує значення χ^2, одержане за формулою (6), і теоретичне значення $\chi^2_{\text{кр}}$, знайдене за відповідними таблицями. В останніх вказується число ступенів свободи $l = k - 1$, де k – число вибірок (всі вибірки мають однакову довжину). Крім того, вказується так званий рівень значу-

щості α. Найчастіше α встановлюється на рівні 0,05 (5 %) або 0,01 (1 %). Рівень значущості являє собою ту мінімальну ймовірність, починаючи з якої можна визнати подію практично неможливою. Наприклад, рівень значущості $\alpha = 0{,}05$ (див. табл. 3). Це означає, що у середньому в 5 випадках із 100 маємо ризик відхилити правильну гіпотезу про розподіл в. в.

Таблиця 3. **Числові значення χ^2 в залежності від l і α**

l	α		l	α	
	0,1	0,05		0,1	0,05
1	2,71	3,84	10	15,99	18,31
2	4,61	5,99	11	21,06	23,68
3	6,25	7,81	15	22,31	25,00
4	7,78	9,49	19	27,20	30,14
5	9,24	11,07	20	28,41	31,41
6	10,64	12,59	24	33,20	36,42
7	12,02	14,07	25	34,38	37,65
8	13,36	15,51	29	39,09	42,56
9	14,68	16,92	30	40,26	43,77

У нас $l = 10 - 1 = 9$, за табл. 3 при $\alpha = 5$ % (0,05) відповідне $\chi^2_{кр} = 16{,}9$. Обчислене за формулою (6) $\chi^2 = 16{,}2$. Математична статистика стверджує, що при вказаних α і $\chi^2_{кр} > \chi^2$ розходження частот можна вважати випадковим. Для нашого випадку 16,9 > 16,2. Це дозволяє визнати гіпотезу про випадковість відхилення 98, 87,…,104 від їх середнього $\bar{x} = 100$ справедливою. Якщо ж $\chi^2_{кр} < \chi^2$, то висунуту гіпотезу відкидаємо.

Проте треба мати на увазі, що жоден з критеріїв згоди, як і критерій χ^2 ("хі-квадрат"), не дає повної відповіді на питання: "Правильна чи ні дана статистична гіпотеза?" Відповіді на подібні питання носять імовірний характер. Не можна на всі 100 % бути впевненим, що гіпотеза правильна, але можна бути впевненим, що вона не *протирічить* дослідним даним.

Ми розглянули приклад, коли із тексту взяті рівні по об'єму вибірки, які дали ряд частот. З'ясували: коливання цих частот навколо середнього випадкові чи закономірні. За допомогою критерія χ^2 можна розв'язати і інші задачі: в досліді одержані дві частоти одного і того ж самого явища мови у двох сукупностях, вибірки з яких були однакового об'єму. Питається: випадкові чи істотні розходження одержаних у досліді частот?

Приклад. У художньому творі із двох вибірок по 500 слововживань кожна були одержані частоти іменників відповідно 270 і 220. Їх середня $\bar{x}$ = 245. Істотна чи ні різниця між ними? Застосовуємо формулу $\chi^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2}{\bar{x}}$: $\chi^2 = \frac{(270-245)^2}{245} + \frac{(220-245)^2}{245} = 5{,}1$. Нехай $\alpha = 0{,}05$. З таблиці маємо, що $\chi^2_{кр} = 3{,}84$; $\chi^2_{кр} < \chi^2$ (3,84 < 5,1). Тому гіпотезу про випадковість розходження частот доведеться відкинути.

Приклад. Маємо вибірки однакового об'єму із різних джерел, в яких дієприкметник повторився відповідно 83 і 65 разів. Різниця велика. Середнє $\bar{x} = \frac{83+65}{2} = 74$. $\chi^2 = \frac{(83-74)^2}{74} + \frac{(65-74)^2}{74} = 2{,}19$. Нехай = 0,05. З таблиці маємо $\chi^2_{кр} = 3{,}84$; $\chi^2_{кр} > \chi^2$ (3,84 > 2,19). У нас немає підстави відкидати гіпотезу про рівність частот, тобто розкид частот носить випадковий характер.

Як застосувати "критерій χ^2", коли вибірки різних об'ємів? Наприклад, одна – об'ємом у 530 слів, а інша – 970 слів. Нехай у них частоти прикметників відповідно 75 і 100. Робимо так:

1) сумуємо частоти: 75 + 100 = 175;

2) сумуємо вибірки: 530 + 970 = 1500;

3) знаходимо відносну частоту: $\frac{175}{1500} = 0{,}116$;

4) 0,116·530 = 61,5;

5) 0,116·970 = 113,5

Тоді $\chi^2 = \frac{(75-61{,}5)^2}{61{,}5} + \frac{(100-113{,}5)^2}{113{,}5} = 4{,}57$.

Нехай $\alpha = 0{,}05$, тоді $\chi^2_{кр} = 3{,}84$; $\chi^2_{кр} < \chi^2$ (3,84 < 4,57) і гіпотеза про несуттєвість, випадковість розходження частот відпадає.

Отже, були розглянуті два типи задач, у розв'язанні яких доцільно було застосовувати критерій згоди "хі – квадрат": 1) узагальнена оцінка величини і характеру коливань частот в їх ряду; 2) оцінка величини розходжень двох частот. Першу задачу можна розв'язати ще за допомогою так званого *коефіцієнта варіації*, а друга заміняється схожою задачею *порівняння часток*.

Коефіцієнт варіації визначається за формулою

$$V = \frac{\sigma}{\bar{x}} \cdot 100\ \%,$$

тобто це є виражене у процентах відношення середньоквадратичного відхилення

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \bar{x})^2}{k}}$$

до середньої частоти $\bar{x}$. Оскільки σ характеризує усереднено сукупність відхилень вибіркових частот від їх середньої частоти, то коефіцієнт варіації, як відношення σ до $\bar{x}$, є задовільною мірою коливальності частот, яка застосовується статистиками при розв'язанні ряду практичних задач. Розглянемо ще раз першу задачу цього розділу:

$$\sigma = \sqrt{\frac{(98-100)^2 + (87-100)^2 + ... + (104-100)^2}{10}} \approx 12,7.$$

Звідси $V = \frac{12,7}{100} \cdot 100\ \% = 12,7\ \%$.

Такий коефіцієнт варіації визнається цілком допустимим для гіпотези про випадковість варіювання частот. Взагалі, якщо $V \le 40\%$, то варіювання частот визнається випадковим. Звичайно, це досить наближено взята границя і там, де треба більша степінь точності, краще застосовувати "критерій χ^2".

4. Порівняння часток

Частка – це відношення спостережувальної частоти до об'єму вибірки. Формула частки:

$$p = \frac{m}{n},$$

де m – частота, n – об'єм вибірки.

Частка ще називається відносною частотою.

Наприклад, якщо у вибірці з 1000 слововживань дієслово зустрічається 250 разів, то частка $p = \frac{250}{1000} = 0,25$ (25 %).

Частки, очевидно, коливаються, як і частоти, коло деякої середньої величини, виражаючи дію закону ймовірності. Якщо у даних

умовах коливання часток підпорядковані одному і тому ж статистичному закону, то це дозволяє обчислити *квадратичне* відхилення M частки, яке визначається за формулою

$$M = \sqrt{\frac{p \cdot q}{n}} \tag{7}$$

$(q = 1 - p)$.

Формула (7) застосовується для порівняння часток одного і того ж явища у двох різних статистичних сукупностях фактів. Наприклад, порівнюють частки присудків у художній прозі Стельмаха і публіцистиці Яворівського.

Для порівняння двох часток формула (7) набуває вигляд

$$E_{1,2} = \sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \tag{3}$$

де $E_{1,2}$ – величина квадратичного відхилення середньої частки двох порівнювальних сукупностей; $\bar{p}$ і $\bar{q}$ – середні частки явищ, що вивчаються; n_1 і n_2 – об'єми вибірок.

Приклад. Є дві вибірки по 1000 слів кожна. В першій 200 дієслів, у другій – 150. Чи можливо припустити гіпотезу про статистичну рівність часток дієслів першої і другої вибірок, тобто чи можна припустити, що фактична розбіжність часток пояснюється законами статистичного варіювання одної і тої ж частки (ймовірності)?

Розв'язання:

$$p_1 = \frac{200}{1000} = 0{,}2; \quad p_2 = \frac{150}{1000} = 0{,}15; \quad \bar{p} = \frac{0{,}2 + 0{,}15}{2} = 0{,}175;$$

$$\bar{p} = 1 - \bar{p} = 0{,}825; \quad E_{1,2} = \sqrt{0{,}175 \cdot 0{,}825\left(\frac{1}{1000} + \frac{1}{1000}\right)} = 0{,}017.$$

Одержане числове значення треба порівняти з різницею часток: $p_1 - p_2 = 0{,}2 - 0{,}15 = 0{,}05$.

Якщо $3E_{1,2} \le |p_1 - p_2|$, то ми можемо відкинути гіпотезу про випадковість, тобто гіпотезу про незначну розбіжність часток. У нас

$3 \cdot 0{,}017 = 0{,}051$; $0{,}015 < 0{,}05$. У нашому випадку ми відкидаємо гіпотезу про випадковий характер розбіжності часток.

Якщо ж $3E_{1,2}$ значно перевищує $|p_1 - p_2|$, то гіпотезу про випадковість неспівпадання часток можна зберегти.

Взагалі, зручніше порівнювати не вибіркові частоти, а частки, наприклад, при порівнянні часток дієслів у стилях: науковому і художньому, у Т. Шевченка і Лесі Українки, у мовах: англійській і хінді, в українській мові XVIII і XX століть.

5. Порівняння середніх вибіркових частот і частотних рядів

Окрім порівняння спостережуваних вибіркових частот і часток лінгвіст може бути зацікавленим і в порівнянні середніх вибіркових частот. Ця задача виникає при дослідженні різних текстів і видів мовлення, різних мовних стилів. Саме тут зручно характеризувати їх середніми частотами та співвідношеннями цих частот.

Приклад. Із текстів письменників А і В взяли по 10 вибірок, кожна по 500 слів. Одержали статистичний ряд розподілу прикметників, наведений у табл. 4.

Таблиця 4. **Значення x_i**

Письменники	i									
	1	2	3	4	5	6	7	8	9	10
А	72	65	78	71	70	74	80	90	68	82
В	80	93	84	83	78	67	85	86	75	89

Статистична задача порівняння двох середніх частот розв'язується, наприклад, за допомогою так званого *квадратичного відхилення* їх різниці згідно формули

$$E_{1,2} = \sqrt{\frac{\sigma_1^2}{k_1} + \frac{\sigma_1^2}{k_2}},$$

де σ_1^2 *і* σ_2^2 – дисперсії двох вибірок; k_1 і k_2 – число вибірок.

Нагадаємо, що дисперсія

$$\sigma^2 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2}{k}.$$

Одержана величина $E_{1,2}$ порівнюється з різницею двох середніх частот, і якщо $3E_{1,2} \le |\bar{x}_1 - \bar{x}_2|$, то гіпотеза про несуттєвість розходжень частот відкидається; в іншому разі – підтверджується.

Для нашого випадку $\bar{x}_1 = 75$; $\bar{x}_2 = 82$.

$$\sigma_1^2 = \frac{(72-75)^2 + ... + (82-75)^2}{10} = 50,8;$$

$$\sigma_2^2 = \frac{(80-82)^2 + ... + (89-82)^2}{10} = 49,4;$$

$$E_{1,2} = \sqrt{\frac{50,8}{10} + \frac{49,4}{10}} = 3,17; \quad 3 \cdot 3,17 = 9,51 > 82 - 75 = 7,$$

тобто середні відрізняються одна від одної несуттєво, вони розійшлися внаслідок звичайного статистичного варіювання одної і тої ж імовірності.

Зауважимо, для порівняння середніх застосовується і *критерій Стьюдента*, який ми тут не розглядаємо.

Отже, лінгвіст, користуючись порівняно невеликим арсеналом статистичних формул, може розв'язати задачу про порівняння *спостережувальних* частот і їх часток. Результати такого порівняння дають відповіді на питання: чи можна спостережені розходження частот або часток пояснити дією однієї і тої ж статистичної, ймовірної закономірності, її випадковим варіюванням, чи ці розходження треба пояснити дією двох різних імовірнісних законів. Якщо підтверджується перше, то говорять про однорідність у співвідношенні факторів, що вивчаються; у іншому випадку, якщо він не підтверджується, – про неоднорідність по відношенню до фактів, що статистично вивчаються.

6. Похибки спостережень і визначення об'єму вибірок із тексту

Математична статистика дає в руки дослідника особливі інструменти, які дозволяють знайти так звану "похибку спостереження", тобто границі, в яких знаходиться "дійсна середня частота" або "дійсна частка", якщо вважати, що ділянка текста, яка не вивчається, однорідна з тією, що вивчається.

Нагадаємо, що інтервал, в якому знаходиться "дійсна середня частота" є ($\bar{x}-\Delta$, $\bar{x}+\Delta$), де $\Delta=\frac{t\sigma}{\sqrt{k}}$, t – особливий коефіцієнт, що залежить від числа вибірок і підраховується за таблицею. Крім того він залежить від надійної імовірності β. Якщо, наприклад, $\beta = 0{,}95$, то це означає, що тільки в 5 випадках із 100 середнє значення частоти виходить за вказаний інтервал. Приклад на застосування цього правила було дано раніше. Але крім абсолютної похибки, що вказує на число одиниць, на які дійсна середня частота може бути більше чи менше вибіркової середньої, вводиться поняття відносної похибки як відношення абсолютної похибки до вибіркової середньої частоти, виражену у процентах або у вигляді десяткового дробу. Введення другого поняття зв'язане з такими міркуваннями: нехай абсолютна похибка $\Delta = 25$ для прикметників при $\bar{x} = 50$ і при $\bar{x} = 500$. Ясно, що у першому випадку вона дуже велика, у другому – мала.

Формула для відносної похибки:

$$\delta=\frac{\Delta}{\bar{x}}=\frac{t\sigma}{\bar{x}\sqrt{k}}.$$

У першому випадку $\delta_1=\frac{25}{50}=0{,}5$ (50 %). У другому випадку $\delta_2 = \frac{25}{500}=0{,}05$ (5 %).

Якщо у досліді вивчаються не середні частоти, а частки і вимагається визначити, яку похибку ми допускаємо у визначенні "дійсної долі" фактів, що вивчаються, то застосовується формула

$$L_p=2\sqrt{\frac{pq}{n}},$$

де 2 – постійний коефіцієнт; p і q – вибіркові частки фактів, що вивчаються, і всіх інших; n – об'єм вибірки.

Наприклад, об'єм вибірки $n = 10000$ слововживань, число іменників в ній – 3500. Їх частка $p=\frac{3500}{10000}=0{,}35$; $q = 1 – 0{,}35 = 0{,}65$.

Тоді $L_p=2\sqrt{\frac{0{,}35\cdot 0{,}65}{10000}}\approx 0{,}01$.

Це значить, що дійсна частка знаходиться в інтервалі ($p - L_p$; $p + L_p$) = (0,35 – 0,01;0,35 + 0,01) = (0,34; 0,36).

Надійність такої відповіді приблизно 95 %. Відносна похибка частки

$$\delta = \frac{L_p}{p} = \frac{2}{p}\sqrt{\frac{pq}{n}} = 2\sqrt{\frac{q}{np}}.$$

У нашому випадку

$$\delta_p = 2\sqrt{\frac{0,65}{10000 \cdot 0,35}} \approx 0,027 \ \ (2,7\,\%).$$

У лінгвістиці відносна похибка у межах 5–10 % вважається допустимою. Формули абсолютної і відносної похибки середньої частоти і частки дозволяють планувати статистичний дослід і визначити або число вибірок k, або, наприклад, об'єм n вибірки.

Із формули $\delta = \frac{t\sigma}{\bar{x}\sqrt{k}}$ маємо

$$k = \frac{t^2\sigma^2}{\delta^2 \bar{x}^2}.$$

На практиці для надійності у 95 % $t \approx 2$, і тоді

$$k = \frac{4\sigma^2}{\delta^2 \bar{x}^2}.$$

Наприклад, σ = 16,5; δ = 0,05; $\bar{x}$ = 90; тоді

$$k = \frac{4 \cdot 16,5^2}{0,05^2 \cdot 90^2} \approx 54.$$

Як бачимо, треба мати велику серію вибірок. Але якщо середнє квадратичне відхилення не 16,5, а 7,5 (що часто зустрічається на практиці мовних явищ), то

$$k = \frac{4 \cdot 7,5^2}{0,05^2 \cdot 90^2} \approx 11.$$

Число вибірок зменшилось. Якщо ж середньоквадратичне відхилення залишається 16,5, то для зменшення числа вибірок збільшуємо δ, наприклад, візьмемо $\delta = 0{,}1$, тоді

$$k = \frac{4 \cdot 16{,}5^2}{0{,}1^2 \cdot 90^2} \approx 13.$$

Перетворюючи формулу визначення абсолютної і відносної похибки частки, можна одержати формули для визначення об'єму вибірки.

Якщо $L_p = 2\sqrt{\frac{pq}{n}}$, то $n = \frac{4pq}{L_p^2}$.

Якщо $\delta_p = 2\sqrt{\frac{q}{np}}$, то $n = \frac{4q}{\delta_p^2 p}$.

Приклад 1. Відомо, що частка прислівників у художній прозі $p = 0{,}07$. Якого об'єму вибірку треба взяти, щоб абсолютна похибка не перевищувала 0,005?

Розв'язання: $n = \frac{4 \cdot 0{,}07 \cdot 0{,}93}{0{,}005^2} \approx 10416.$

Приклад 2. Частка прислівників $p = 0{,}07$. Якого об'єму вибірку треба взяти, щоб забезпечити відносну похибку у 0,05?

Розв'язання: $n = \frac{4 \cdot 0{,}93}{0{,}05^2 \cdot 0{,}07} \approx 21257.$

На практиці n беруть або 10500, або 21000. Досвід застосування статистики для вивчення основ морфології і синтаксису в різних стилях, наприклад, російської літературної мови XIX–XX століть переконує у тому, що 10 або 20 вибірок довжиною у 500 слововживань кожна, дають досить задовільну точність як середніх частот, так і часток.

IV. КОРЕЛЯЦІЯ І РЕГРЕСІЯ

Зв'язок між ознаками може бути функціональним і кореляційним (статистичним).

Функціональним називають зв'язок між ознаками, коли кожному значенню однієї змінної (аргументу) за певним правилом ставиться у відповідність єдине значення другої змінної (функції).

Такі зв'язки спостерігаються в математиці (площа круга $S = \pi R^2$), фізиці (сила струму $I = U/R$), астрономії і навіть у лінгвістиці (погодження за родом і числом). Але у соціально-економічних, лінгвістичних дослідженнях функціональна залежність зустрічається рідко. Тут частіше зустрічається залежність, за якою кожному значенню однієї величини може відповідати декілька значень іншої, причому ці значення самі по собі є, як правило, випадковими величинами з певним законом розподілу. Такі залежності між ознаками одержали назву *кореляційних* (*статистичних*) *залежностей*.

Наприклад, на 10 однакових ділянках поля внесли однакову кількість добрив. Але врожай зернових одержали різний, так як на нього впливають й інші фактори (якість ґрунту, кількість опадів і т.п.). У цьому випадку знання поведінки однієї ознаки (кількість внесених добрив) дає можливість прогнозувати поведінку іншої ознаки (врожайність) лише з певною долею ймовірності.

Лінійна кореляція припускає лінійну залежність: при зростанні (спаданні) однієї випадкової величини інша випадкова величина зростає (спадає).

Така тенденція може бути вираженою сильно і слабо. В цьому випадку говорять про *тісноту статистичної залежності*. Як правило, характер і тіснота зв'язку визначається *коефіцієнтом кореляції*, величина якого коливається від –1 до +1. Від'ємні значення коефіцієнта кореляції свідчать про те, що із зростанням (спаданням) однієї випадкової величини друга проявляє тенденцію до спадання (зростання). Додатна кореляція говорить про те, що із зростанням (спаданням) однієї випадкової величини друга також зростає (спадає). Чисельне значення коефіцієнта кореляції вказує на тісноту статистичного зв'язку. Причому, чим ближче за модулем коефіцієнт кореляції до 1, тим цей зв'язок ближче до функціонального, а при наближені до 0 – вказує про слабкий зв'язок або про практичну його відсутність.

Приклад. В кожній із 15 вибірок, взятих із авторської художньої прози Герцена (кожна вибірка по 500 слововживань), опинилося по x іменників і y займенників.

X і Y – випадкові величини. Встановити статистичну (кореляційну) залежність між ними. Дані зведені у табл. 5 (числові дані округлені до десятих).

Таблиця 5. **Числові дані**

№	x	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	y	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	171	–7,3	52,9	10	3,5	12,5	–25,7
2	187	8,7	76,2	98	1,5	2,3	13,4
3	164	–14,3	203,6	114	17,5	307,3	–250,2
4	186	7,7	59,8	85	–11,5	131,6	–88,7
5	181	2,7	7,5	68	–28,5	810,5	–77,7
6	168	–10,3	105,5	98	1,5	2,3	15,7
7	201	22,7	516,7	92	–4,5	20,0	–101,5
8	150	–28,3	799,2	109	12,5	157,0	–354,2
9	183	4,7	22,4	105	8,5	72,8	40,4
10	169	–9,3	85,9	109	12,5	157,0	–116,2
11	158	–20,3	410,9	106	9,5	90,8	–193,2
12	183	4,7	22,4	94	7,5	56,7	35,6
13	170	–8,3	68,4	102	5,5	30,6	–45,7
14	213	34,5	1206,2	71	–25,5	648,7	–882,6
15	190	11,7	137,6	96	-0,5	0,2	–5,5
Σ	2674		3775,2	1447		2500,3	–2036,1

За цими даними підраховуємо середні значення $\bar{x}$ і $\bar{y}$:

$$\bar{x} = \frac{2674}{15} = 178,27 \approx 178,3; \quad \bar{y} = \frac{1447}{15} = 96,47 \approx 96,5.$$

Тісноту зв'язку встановлює коефіцієнт кореляції, позначений через r_{xy}, що визначається формулою

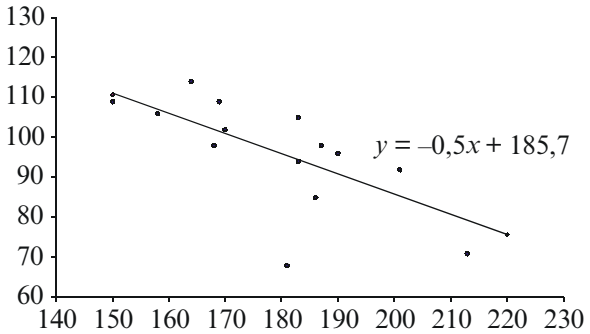
$$r_{xy} = \frac{\sum_{i=1}^{k}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{k}(x_i - \bar{x})^2 \sum_{i=1}^{k}(y_i - \bar{y})^2}}.$$

Для нашого випадку $r_{xy} = -\dfrac{2036,1}{\sqrt{3775,2 \cdot 2500,3}} = -0,7$.

Зв'язок достатньо тісний і з ростом x (іменників) маємо тенденцію до зменшення y (займенників), на що вказує знак "мінус".

За розташуванням точок (x_i, y_i) (див. рисунок) видно, що залежність між x і y лінійна, отже,

$$y = ax + b. \tag{8}$$



Рівняння (8) називається *рівнянням прямої лінії регресії* y на x. Доводиться, що

$$a = \frac{\sum_{i=1}^{k}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{k}(x_i - \bar{x})^2}, \quad b = \bar{y} - a\bar{x}.$$

Для нашого випадку $a = \dfrac{-2036,1}{3775,2} = -0,5$, b = 96,5 + 0,5·178,3 = = 185,7. Отже, $y = -0,5x + 185,7$.

Будуємо дану пряму по двом точкам. Як бачимо, точки (x_i, y_i) розташовані досить близько від даної прямої.

Приклад для самостійної роботи

У 15 вибірках рівного об'єму із деякого твору виявилось по x прикметників і y дієприкметників. X і Y – випадкові величини. Вста-

новити кореляційну залежність між ними. Вихідні дані задаються табл. 6.

Таблиця 6. **Вихідні дані**

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X	49	53	58	39	37	48	43	56	38	37	39	46	48	37	45
Y	15	10	6	19	16	12	12	7	17	17	18	13	14	17	11

Відповідь: $r_{xy} = -0{,}9$; $a = -0{,}5$; $b = 36$.

РЕКОМЕНДОВАНА ЛІТЕРАТУРА

1. *Головин Б.Н.* Из курса лекций по лингвистической статистике. – Горький: ГГУ, 1966. – С. 3–94.

2. *Гурский Е.И.* Теория вероятностей с элементами математической статистики. – М.: Высшая школа, 1971. – С. 268–318.

3. *Лесохин М. М., Лукьяненков К.Ф., Пиотровский Р.Г.* Введение в математическую статистику. – Минск: Наука и техника, 1982. – С. 7–220.

4. *Носенко Н.А.* Начала статистики для лингвистов. – М.: Высшая школа, 1981. – С. 3–154.

5. *Перебийніс В.І.* Статистичні методи для лінгвістів: Посібник. – Вінниця: Нова книга, 2002. – С. 3–170.

6. *Толбатов Ю.А.* Математична статистика та задачі оптимізації в алгоритмах і програмах. – К.: Вища школа, 1994. – С. 88–218.